

Discrimination of performance tiers and prediction of success in introductory physics courses using a statistical method for establishing cutoff scores

Maher M. Alarfaj

King Faisal University, Saudi Arabia

Charles Secolsky

County College of Morris, Randolph, New Jersey

Fahad S. Alshaya

King Saud University, Saudi Arabia

The Excellence Research Center of Science and Mathematics Education, Saudi Arabia.

Abstract

This study sheds light on the prediction of success using cutoff scores for student grades adopted for a required Physics pathway course for study in a health professions program at King Saud University in Saudi Arabia. Data on course grade and GPA for approximately 10,000 students enrolled in this course between 2008–2014, were analyzed. Receiver Operating Characteristic (ROC) curve analysis was used to determine cutoffs for course grades using ranges of GPA. This procedure has promise as a new method for quantitatively arriving at cutoff scores using an external criterion requiring less human judgment than most existing standard setting methods. The cutoff scores produced show that GPAs of students who complete the Physics course yield successive performance tiers that are lower than expected. In addition, the correlation between GPA and course grade for Physics is only 0.63 and therefore only 39% of the variation in GPA explains course grade. As a result of the findings of the study, the decision was made to maintain the existing standards thereby requiring higher grades in the Physics course for students seeking to enter a health professions course of study.

تسلط الدراسة الحالية الضوء على الدرجات الحدية وقدرتها التنبؤية على نجاح الطلبة في مقرر الفيزياء الاولي (فيز 145)، والذي يعد متطلبا اساسيا على طلبة التخصصات الصحية في جامعة الملك سعود بالمملكة العربية السعودية، وعلية فقد تم جمع البيانات المتعلقة بدرجات هذا المقرر لما يقارب من 10000 من الطلبة اللذين التحقوا بهذا المقرر بين الاعوام 2008 - 2014 م، ومعدلاتهم التراكمية. وتحليل هذه البيانات، فقد تم استخدام منحني خاصية تشغيل العمليات (ROC) Receiver Operating Characteristic لتحديد الدرجات الحدية لنطاقات مختلفة من المعدلات التراكمية، حيث تعد هذه الطريقة من الطرق الكمية الحديثة للوصول الى الدرجات الحدية والحد من التأثير البشري. وعلية فان الدرجات الحديه التي تم الحصول عليها قد اشارت الى ان المعدلات التراكمية لمن اجتاز هذا المقرر كان ضمن نطاقات اقل من المتوقع، كما ان العلاقة الارتباطية بين المعدلات التراكمية ودرجات الطلبة بلغت 0.63 ، مما يعني ان 39% من التباينات في المعدلات التراكمية للطلبة يمكن ان تسهم في تفسير درجاتهم في ذلك المقرر. ووفقا لما تم الحصول عليه من نتائج يتبين اهمية المحافظة على معايير المقرر الحالية مع تأكيد حصول الطلبة الراغبين للالتحاق بالتخصصات الصحية على درجات مرتفعة في مقرر الفيزياء.

Introduction

Grading of students' performance in higher education typically involves reference to cutoff scores which define bands of performance such as "Fail", "Excellent", or letter grades (C+, B-, etc.). Most procedures

for arriving at cutoff scores for student grades require expert judgment of item difficulty and examinee ability. Introduced in this article is a new method for establishing cutoff scores, based on the statistical relationship between (a) existing grades in an introductory Physics course used for accepting students for study in the health professions and (b) the variable undergraduate grade point average (GPA). The statistical procedure used to determine each grade level or tier was logistic regression followed by Receiver Operating Characteristic (ROC) curve analysis. The contribution of this article is that the proposed procedure relies less on human judgment than many existing standard setting methods.

Introductory college courses represent a gateway for many future educational decisions; it is therefore very important that the grading process in these courses be defensible. Letter grades are typically used to indicate a student's progression through the grading system, and as such, are used to differentiate students based on performance. Hence, they are thought to be a meaningful reflection of student academic potential. Yet Guskey and Anderman (2013) have pointed out a primary weakness in grading systems based mostly on cutoff scores with borderline performance tiers. It is that there exists a wide variety of perceptions among educators as to what constitutes mastery of a particular topic or course. Belfield and Crosta (2012) collected data from several community colleges across the U.S. and calculated the accuracy rates and four validity metrics for placement tests based on cutoff scores. They found high error rates using these placement cutoffs. The severe error rate for English was 27 to 33 percent, meaning that on average three out of every ten students are mis-assigned. Given these high error rates, it made sense to explore the use of a procedure that is based less on perceptions of what constitutes performance mastery and more on statistical objectivity.

At the college level, placement tests are widely used to discriminate various performance tiers that bear high risks of inadvertently assigning students into developmental or remedial coursework. Some authors have used logistic regression analysis and ROC curve analysis to make judgments of the validity of cutoff scores for placement tests. Cutoff scores are set up to discriminate between student performance tiers, usually denoted by letter grades aligned with cutoff scores. Grade point averages (GPAs) and course grade composites are effectively influenced by preset cutoff scores. Therefore, cutoff scores play a critical role in classifying students into a scale composed of contiguous levels for assigning a level of performance to student coursework. It has been recommended that cutoff score decisions be reexamined at least every five to seven years, unless performance issues arise suggesting that such review be done sooner (Morgan & Michaelides, 2005). Consequently, this article is responding to this call for greater understanding of cutoff scores by shedding light on how cutoff scores are functioning in an introductory physics course.

Statement of the problem

Introductory physics courses at King Saud University are offered by the Physics Department in the College of Science, and are required by various colleges as part of first year requirements. It is especially important to remember that for many students, such courses are considered as indicators of their abilities to pursue a field of study. In addition, it is vital to monitor the initial performance standards and to identify corresponding cutoff scores to enable accurate categorization of students' performances (Kane, 1994). Therefore, this study aimed to answer the following questions:

1. What is the effectiveness of a new quantitative procedure for establishing cutoff scores based on logistic regression and ROC curve analysis?
2. Do administered cutoff scores represent a significant indication of the initial performance tiers on which subsequent physics course grades are based?

Background

Health science colleges, in Saudi Arabia, have become a desirable choice for high school graduates. Over the past decade, the admission at these schools has increased considerably in response to the high demand. For example, the number of enrolled students in health science colleges increased from a total of 2,712 students on record in 2001 to a total of 23,293 students on record for the study period of this study, (2008-2014) (Ministry of Education, 2016). Admission to health profession programs at Saudi Arabian universities comprises two stages. The first stage is meeting the qualifications to enter the medical track; this depends solely on high school performance and scores from two standardized tests (the Standard Achievement Admission Test and the General Aptitude Test – both locally written and administered by the National Center for Assessment in Higher Education). The second stage requires passing the preparatory program, which extends for one academic year and offers courses primarily in English, mathematics, chemistry, biology, and physics.

Generally, introductory physics courses at King Saud University are offered by various colleges as part of the first-year requirements. Among these courses, there is *Physics 145*, offered to students intending to take a medical or health pathway. It is three credit hours, which includes two hours for lectures and one hour for labs. It is a competitive course enrolling many of the best and brightest students. In addition, this physics course inspires the mind towards logical thinking and deduction, which is vital for a medical professional. It includes instructional classes and laboratory sessions focusing on various physics concepts and medical applications. The classes include topics in mechanics, electricity, mirrors and lenses, light, nuclear physics, and ionizing radiation. The lab for this course provides a hands-on experience, with students performing experiments, recording notes, responding to pre- and post-experiment questions and writing a final lab report. The experiments cover the simple pendulum, Hooke's law, friction, vector components, viscosity, Ohm's law, lenses, and free fall Newton's laws.

As a dominant practice, letter grades are used to track student performance and progression in this course using an A-F scale. In every official course, students are assigned grades according to the following grading system:

- 95 to 100 is A +.
- 90 to less than 95 is A.
- 85 to less than 90 is B+.
- 80 to less than 85 is B.
- 75 to less than 80 is C +.
- 70 to less than 75 is C.
- 65 to less than 70 is D+.
- 60 to less than 65 is D.
- Less than 60 is F.

The overall academic success of a student at graduation can be characterized in two ways. One is the final cumulative GPA. The GPA for a student is typically computed by converting the letter grade for each course to date into a numerical value (using a scale of 1(F) to 5(A)), multiplying this numerical value by the number of credits allotted to the course, calculating the total of these weighted values and then dividing by the student's total number of credits. The GPA may then be categorized into bands as shown below.

- "Excellent" for GPAs 4.50 and above.

- “Very Good” for GPAs from 3.75 to less than 4.50.
- “Good” for GPAs from 2.75 to less than 3.75.
- “Satisfactory” for GPAs from 2.00 to less than 2.75.
- “Fail” for GPAs less than 2.

Use of these two widespread categorical grading scales likely stemmed from their accessibility in communication with students; however, they are subjective, and too abstract to be used alone to reflect student performance (Marzano, 2010). It is possible that various institutions and teachers offering the same course share the same grade to indicate student performance, but use divergent weighted tasks to arrive at a grade. In higher education, the course grading system is apparently tied to an administered test. Alqataee (2012) claimed that teachers always deviated from the test construction rules by implicitly overemphasizing certain criteria. In addition, Stiggins and Bridgeford (1985) stressed that teacher-made tests lack quality control and tend to favor domains other than those intended. These factors affect the quality of testing and thus the quality of grading. However, in introductory physics courses, including *Physics 145*, it is a widely-held practice that faculty who teach the same course, construct the exam, and check compliance with the same test specifications.

The quality of constructed tests and administration of the tests according to standards can have an impact on a grading system, which could be potentially controlled and managed. For instance, if the standard is to apply mathematical knowledge in multiple authentic ways, the grading system will be orchestrated to weight students’ performance according to this multiplicity. However, the grading system would still have a basic weakness, namely supporting arbitrary and uncertain borderlines or cutoff scores on the 0-100 scale. Guskey and Anderman (2013) stated that percentage cutoffs for any scale are a subjective decision, and they imply little about the objective evaluation of a student’s performance. A more objectively derived cutoff score would enable the determination of student performance tiers. The logic behind creating these cutoff points is a dominant practice used to sort out students based on customary practice (Pitoniak & Morgan, 2012). Cutoff scores can be set using the Contrasting Groups method (Livingston & Zieky, 1982). This method uses judgments about test-takers with respect to whether teachers believe students will succeed on some external measure. Using this method, two distributions of GPA are formed based on this external measure: one of examinees deemed by their teachers to have a high enough GPA to be able to pass *Physics 145* and one distribution of examinees deemed by their teachers as not having a high enough GPA for success in *Physics 145* (the gateway course). In the present study, instead of just considering whether students had high enough or too low GPAs to succeed in *Physics 145* (the pass/fail cutoff score for the course in general), cutoff scores separating different passing *Physics 145* course grades were created from corresponding ranges of GPA. Logistic regression is a fairly robust approach for classifying students with respect to some established standard – or in this study a range – or between two such ranges for finding a cutoff score where the dependent variable, course grade, is dichotomous, meaning that the dependent variable takes on only two values. In the present study, the dichotomy consisted of two ranges, with each of the two ranges surrounding a middle range within which the cutoff was sought (Morgan & Michaelides, 2005; Secolsky et al., 2013). There are a number of standard setting methods, including the modified Angoff method, the Contrasting Groups approach, the borderline group approach (Livingston and Zieky, 1982), and the bookmark approach (Mitzel, Lewis, Patz, and Green, 2001) to name a few; however, logistic regression is an essential and concise approach for standard-setting and validation. It is more quantitative in nature and less directly reliant on expert judgment than other approaches (see Pitoniak & Morgan (2012) for a description of various standard setting procedures).

Method

This study was conducted on one of the introductory Physics courses at King Saud University to provide an example of how to statistically arrive at cutoff scores for different performance tiers. The course chosen was *Physics 145*, a popular course for those students seeking to enter the health professions as well as other disciplines. Typically, setting cutoff scores requires some use of judgment by the researcher; for example, the Contrasting Groups method (Livingston & Zieky, 1982) necessitates selecting another variable or measure on which to classify students into different tiers. This was the situation in the present investigation; thus, the judgments in this study resided in the selection of the ranges on which to separate groups as a function of performance tiers (the ranges of GPA needed to create cutoff scores for *Physics 145* course grades). Once these ranges were selected, the remainder of the procedure was mostly statistical. Administered cutoff scores can be judged by whether they represent a significant indication of the initial performance tiers on which subsequent physics course grades are based.

Descriptive statistics

There were 10,795 records stored electronically of students who had initially enrolled in *Physics 145* in the academic years 2008-2014. Of this group, 837 students (7.8%) had missing course grades. The variable used to create cutoff scores was GPA, excluding the current *Physics 145* grade; a histogram of GPAs for the valid 9,858 students is shown in Figure 1. Since all valid 9,858 student records were eligible for inclusion in the analysis, no sampling method was conducted. This allowed more information to be used than by using random, stratified or systematic sampling, and it was not necessary to make inferences to a population of observations (see Sudman, 1976). Observations were treated as missing listwise if the student record contained a blank value for GPA and/or course grade. The statistical software package used for the analyses in this study was SPSS Statistics Version 23.0 (IBM, 2015).

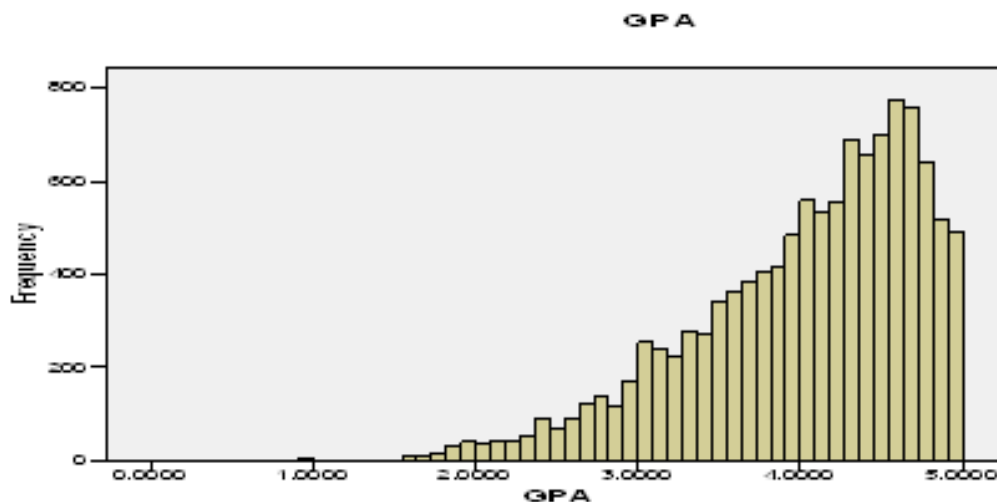


Figure 1: GPA distribution.

As can be seen from the histogram, the GPA distribution is somewhat skewed to the left: there exists a preponderance of GPAs between 4.0 and 5.0, as would be expected for students planning to enter the

health professions. The distribution for the variable course grade was also somewhat negatively skewed. Other descriptive statistics for the two variables used for creating cutoff scores for course grades appear in Table 1.

Table 1: Descriptive statistics for Course Grade and GPA.

	Course Grade	GPA
N	10795	11632
Valid Missing	837	0
Mean	71.263270	4.022082
Median	72.00000	4.180000
Std. Deviation	16.062655	.7006914
Skewness	- 0.847	- 0.911
Kurtosis	1.198	0.430
Minimum	2.0000	0.50000
Maximum	100.0000	5.0000

Receiver Operating Characteristic (ROC) curve analysis was used to determine cutoffs for course grade using ranges of GPA. ROC is a logistic regression based procedure.

Receiver Operating Characteristic (ROC) curve analyses (B+ to A)

Presented below are the procedures for the calculation of the cutoff point for course grade between B+ and A using the selected ranges of GPA. The ranges of GPA used were 4.00-4.24 and 4.75-4.99. GPA was on a scale ranging from 0 to 5.0. Note that there is a gap between these two ranges of GPA, which extends from 4.25 to 4.74. This gap in the ranges can be considered associated with a course grade of A-. Since each set of two ranges used in the analysis were not adjacent, as we skipped over one range, we sought to obtain a cutoff score for the range that was ‘in between’ the two non-adjacent ranges.

For the ROC analysis used in this example, GPAs in the ranges 4.75-4.99 and 4.00-4.24 were treated as a dichotomized variable, such that GPAs ranging from 4.00 through 4.24 were recoded to ‘0’ and those in the range 4.75 thru 4.99 were recoded to ‘1’. This dichotomized variable is the ‘state’ variable in the ROC analysis. The course grade scores are the ‘test scores’ in this analysis. The Case Processing Summary below (Table 2) indicates that 1,366 students in *Physics 145* had GPAs ranging from 4.75-4.99 while 1,370 students had GPAs ranging from 4.00-4.24.

Table 2: Case processing summary.

L_V_H	Valid N (listwise)
Positive ^a	1366
Negative	1370
Missing	8896

Note: Larger values of the test result variable (s) indicate stronger evidence for a positive actual state.

^a : the positive actual state is 1.00

The ROC curve for this analysis is displayed in Figure 2.

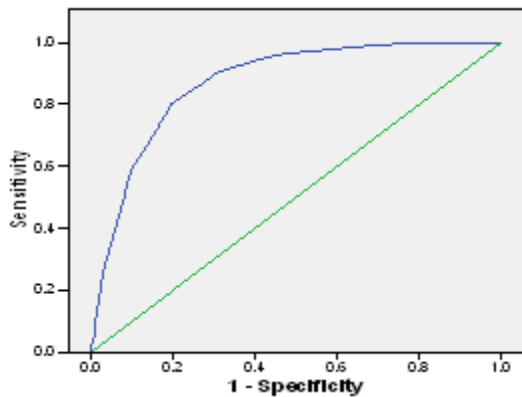


Figure 2: Average sensitivity and specificity for the selected cutoff score using 4.00-4.24 and 4.75-4.99 as the GPAs.

Visual inspection of the ROC suggests that course grade is a good ‘discriminator’ of student GPAs in the two ranges under investigation, i.e., 4.00-4.24 vs.4.75-4.99. A discriminating GPA is one which has a high true positive rate, i.e., high sensitivity, and a low false positive rate (specificity) across a range of course grade cutoff scores. Such screening tests are graphically depicted by ROC curves with a substantial “bow” toward the upper left hand corner of the ROC graph. The fact that this ROC curve does extend toward the upper left hand corner of the graph is one indication that course grade is fairly effective as a discriminator of the two subgroups of GPA corresponding to ‘B+’ vs. ‘A’. It is possible to quantify how quickly the ROC curve rises toward the upper left hand corner of the graph by measuring the *area under the curve* (AUC; see Table 3). A value of ‘1’ for this statistic corresponds to a ‘perfect’ test, one which can accurately discriminate between the two values of the ‘state’ variable, here, B+ vs. A. On the other hand, a value of 0.5 suggests a ‘worthless’ test. Stated somewhat differently, a value of 1 for the AUC statistic indicates that the test is maximally discriminating because it achieves a 100% true positive rate (sensitivity) and a 0% false positive rate (1-specificity). That is to say, it always correctly identifies A students and never incorrectly identifies a B+ student as an A student. Of course, in the ‘real world’, AUC statistics are not typically found at these extremes. While there are different standards for interpreting the magnitude of the AUC statistic, depending upon the field of inquiry, the following ones represent a fairly conservative rule of thumb for the purpose of standard-setting (Rice & Harris, 2005):

- 0.50 – 0.70 ‘low’ discrimination or accuracy
- 0.70 – 0.90 ‘moderate’ discrimination or accuracy
- 0.90 ≥ ‘high’ discrimination or accuracy

As seen in the Table 3, the AUC statistic in this analysis is .872 which is statistically significant and indicates moderately accurate discrimination.

Returning to the plot of the ROC curve, the point at which the maximum curvature occurs corresponds to the optimal trade-off between sensitivity and specificity and for that reason also corresponds to the optimal cutoff score for the test (course grade). While it is possible to at least approximately identify this

point from the ROC curve, it is not possible to determine the test score, i.e., the actual course grade, which corresponds to the point of maximum curvature.

Table 3: Area Under the Curve (AUC): test result variable course grade.

AUC	Std. Error ^a	Asymptotic Sig. ^b	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.872	0.007	0.000	0.859	0.885

Note: The test result variable: course grade has at least one tie between the positive actual state group and the negative actual state group. Because a small number of course grades were associated with both positive and negative actual values of the modeled GPA state variable, some bias may be present.

^a : Under the nonparametric assumption

^b : Null hypothesis: true area = 0.5*

One intuitively appealing way of identifying the optimal cutoff score is to take the average of the sensitivity and specificity scores for each test score (course grade) and choose the test score which has the highest average. The problem with this approach, in this data set, is that there are course grades which have very low sensitivities and very high specificities (or vice versa) but which lead to averages of these two statistics which are among the highest observed. One would not want to use a test score (course grade) with this kind of diagnostic profile. It would be preferable to identify the highest average score derived from sensitivity and specificity values which are themselves as high as possible and as similar to each other as possible. That is to say, one would like a cutoff score for the course grade which exhibits both high sensitivity and high specificity indicating that it accurately discriminates between the test scores which define the two subgroups of the state variable, here B+ vs. A, thereby obtaining the cutoff score for a course grade of A-. In other words, the ranges of GPA used for determining the cutoff score between course grades B+ and A were not adjacent. We thereby found the cutoff score for A- using the respective non-adjacent ranges of GPA for course grades of B+ and A.

In order to do that, the criterion for selecting the optimal cutoff score had to be modified so that it was now the highest average of the grade point averages' sensitivities and specificities subject to two additional constraints. First, both the sensitivity and the specificity values for the optimal cutoff score must be greater than 0.50. The criterion of 0.50 derives from the fact that a cutoff score with a 50% true positive rate (sensitivity) and a 50% false positive rate (1-specificity) is a minimally discriminating test score, i.e., it is no better than flipping a coin as to whether a student with that course grade will be classified into one group or the other. Second, the discrepancy between the sensitivity and specificity values of the optimal cutoff score should be as small as possible subject to the first constraint that both must be greater than .50. Doing so promotes the selection of an optimal cutoff score with a 'good' diagnostic profile, i.e., high sensitivity and high specificity values.

By imposing these criteria on the selection of the optimal cutoff score several desiderata are implemented. Firstly, no 'lopsided' diagnostic profile of sensitivity and specificity values such as .80 and .20 (or vice versa) will determine the optimal cutoff score for discriminating between the two groups (B+ vs. A in this example) simply because it has the highest average value of the sensitivity and specificity values. Secondly, the selected cutoff score is not allowed to have a sensitivity value or a specificity value which is less than minimally acceptable. Thirdly, the selected cutoff score is maximally discriminating because it will have the highest average sensitivity and specificity value subject to the two constraints outlined in the preceding paragraph.

Table 4 lists (up to 20) test scores (course grades) with the highest average values of their respective sensitivity and specificity values subject to the constraints that those sensitivity and specificity values both be greater than .50 and the discrepancy between the sensitivity and specificity values be as small as possible. As seen in Table 4, the course grade with the ‘best’ cutoff score is 80.50. This particular course grade corresponds to the point of maximum curvature in the ROC graph (Figure 2).

Table 4: Sensitivity and specificity values for twenty course grades.

Test_Score	Sensitivity	Specificity	Disc_Sens_Spec	Mn_Sens_Spec
80.50	.8016	.8058	.0042	.8037
81.50	.7657	.8197	.0540	.7927
82.50	.7343	.8343	.1000	.7843
83.50	.7218	.8380	.1161	.7799
84.50	.7189	.8394	.1205	.7792
79.50	.8748	.7204	.1544	.7976
78.50	.8770	.7197	.1573	.7984
77.50	.8792	.7168	.1624	.7980
76.50	.8982	.7015	.1968	.7999
75.50	.9114	.6759	.2355	.7937
85.50	.5944	.9000	.3056	.7472
74.50	.9422	.5971	.3451	.7696
73.50	.9436	.5942	.3495	.7689
72.50	.9466	.5839	.3626	.7653
86.50	.5373	.9139	.3765	.7256
71.50	.9553	.5635	.3918	.7594
70.50	.9649	.5234	.4415	.7441

In Table 4, the values of Test Score refer to 20 different cutoff scores for course grade with corresponding sensitivity and specificity values from the ROC analyses. *Disc Sens-Spec* refers to the discrepancy between sensitivity and specificity values. This indicator was simply obtained by subtracting sensitivity and specificity values for the 20 different cutoff scores. *Mn Sens-Spec* is simply the mean of sensitivity and specificity values.

Figure 3 overlays the test scores (course grades) on the ROC curve. Although the graph is dense, by cross-referencing the original ROC graph above and the overlay ROC graph below, it can be inferred that the point of maximum curvature in both graphs corresponds to a test score (course grade) of 80.50. While this is the “optimal” cutoff score, note that its sensitivity value (.80) and its specificity value (.81) are not particularly high values for either of these two diagnostic statistics, each of which ranges from ‘0’ to ‘1’. This is consistent with the fact that the AUC statistic above (.87) finds that the course grades are only moderately good discriminators of membership in the B+ vs. the A subgroups.

The above ROC analysis was used to find the cutoff score between a letter grade of A and a letter grade of B+. This was done in order to use the dichotomized non-adjacent ranges of GPA: 4.00 - 4.24 and 4.75 - 4.99. The course grade cutoffs as well as their corresponding GPA ranges used to obtain the cutoffs appear in Table 5.

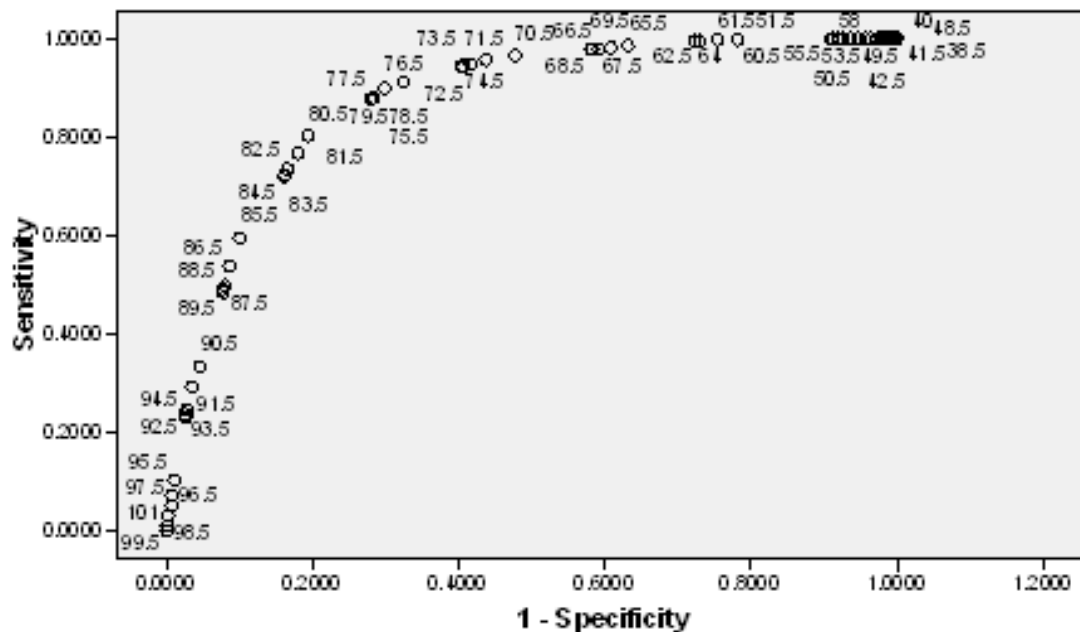


Figure 3: ROC curve with cutoff points.

Results

As can be seen from the example presented, 80.5 is the course grade cutoff score for the dichotomized ranges of GPA: 4.00-4.24 and 4.75-4.99. Table 5 contains the remainder of the cutoff scores for each set of GPA ranges. Note from Table 5 that the cutoff scores steadily decrease as a function of GPA ranges, as can be seen in the logistic regression and ROC curve analyses. In addition, the cutoff score is unexpectedly low in the analysis illustrated above: where 80.5 is the cutoff in question. Overall, the cutoff scores for *Physics 145* tended to be low or moderate, implying that low and moderate course grades are obtained for even students with high GPAs, including GPAs above 4.00.

Table 5: Cutoff Scores for All Course Grades and Corresponding GPAs.

Grade Point Averages	Course Grade Cutoff point
5.00-5.00	94.50
4.50-4.74	
5.00-5.00	94.50
4.25-4.49	
4.75-4.99	80.50
4.00-4.24	
4.50-4.74	74.50
3.75-3.99	
4.25-4.49	69.50
3.50-3.74	
4.00-4.24	65.50
3.25-3.49	

3.75-3.99	63.50
3.00-3.24	
3.50-3.74	60.50
2.75-2.99	
3.25-3.49	59.50
2.00-2.75	
2.00-2.75	42.50
less than 2.00	

Discussion

The establishment of cutoff scores necessarily requires some degree of human judgment, which must be combined with a statistical cutoff score method for arriving at desired cutoff scores. This is the case for any of the methods commonly employed. In the present study, logistic regression and ROC curve analysis were based on the optimization of sensitivity and specificity values for judgmentally arriving at GPA ranges, but the procedure introduced is less dependent on direct human judgment.

The cutoff scores obtained show that the GPAs of students who complete *Physics 145* yielded successive performance tiers that are lower than expected. This might be due to the possibility that other courses that have grades included in the GPA calculation are less difficult than *Physics 145*. Such courses make up the vast majority of course grades used to compute GPAs in this investigation. It is also of interest that the correlation between GPA and course grade for *Physics 145* is only 0.63, implying that only 39% of the variation in GPA explains course grade. Other factors could contribute to explaining course grade, which could be the core of further studies. In addition, the contribution of GPA variation in explaining course grade may differ for other introductory physics courses required for other pathways. This comparison will drive more discussion on the connection between GPAs and introductory physics course grades.

The university was in the position of having to decide whether to make the *Physics 145* gateway course more lenient or else maintain its rigorous standards of being difficult to obtain a high grade. If the existing high standard found in this analysis is maintained by the university for this course, admittance to a graduate program in the health sciences such as a medical school would tend to make for a more selective group of students in health science fields.

As a result, the findings of this study led to the decision to maintain the existing high standards thereby continuing to require difficult to obtain higher course grades in *Physics 145* for admittance to health profession programs at King Saud University. It was recommended that the cutoff scores and related grading practices should remain intact until such a time as a follow-up study is conducted to determine if there is any adverse impact for students receiving somewhat lower grades for this gateway course, seemingly used as a screening mechanism for entrance into health professions. In the future, it would also be important to examine cutoff scores for course grades for other introductory Physics courses and for gateway courses in the course offerings of other departments.

Acknowledgment

This research was conducted as part of the assessment and development initiative by the introductory physics courses group and is supported by the Excellence Research Center of Science and Mathematics Education, King Saud University, Saudi Arabia.

References

- Alqataee, A. (2012). *The effect of the distributions of college course grades on the NCA admission test predictive validity*. Paper presented at the first International Conference on Assessment and Evaluation, 2-4 December (pp, 135- 146). Riyadh: NCAHE.
- Belfield, C., & Crosta, P. (2012). *Predicting success in college: the importance of placement tests and high school transcripts* (CCRC Working Paper No. 42). New York, NY: Columbia University, Teachers College, Community College Research Center.
- Guskey, T. & Anderman, E. (2013). In search of a useful definition of mastery. *Educational Leadership*, 71(4), 18-23.
- IBM Corp. Released 2015. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY. IBM Corp.
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425-461. <http://dx.doi.org/10.3102/00346543064003425>
- Livingston, S., & Zieky, M. (1982). *Manual for setting standards on the basic skills assessment tests*. Princeton, New Jersey: Educational Testing Service.
- Marzano, J. (2010). *Formative assessment and standards-based grading*. Bloomington, IN: Marzano Research Laboratory.
- Ministry of Education. (2016). *A statistical summary of admitted students in 2013*. Riyadh: Higher Education Statistics Center.
- Mitzel, H., Lewis, D., Patz, R., & Green, D. (2001). The bookmark procedure: psychological perspectives. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Morgan, D., & Michaelides, M. (2005). *Setting cut scores for college placement*. Research Report No. 2005-9, New York: The College Board.
- Pitoniak, M. & Morgan, D. (2012). Setting and validating cut scores for tests. In C. Secolsky and D.B. Denison (Eds.) *Handbook on measurement, assessment, and evaluation in higher education* (pp 343-366). New York: Routledge.
- Rice, M., & Harris, G. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior*, 29, 615-620.
- Secolsky, C., Krishnan, K. & Judd, T.P. (2013). Using logistic regression for validating or invalidating initial statewide cut-off scores on basic skills placement tests at the community college level. *Research in Higher Education Journal*, 19, 1-21.
- Stiggins, R., & Bridgeford, N. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22, 271-286.
- Sudman, S. (1976). *Applied sampling*. New York: Academic Press.
- Alarfaj, M., Secolsky, C., & Alshaya, F. (2017). Discrimination of performance tiers and prediction of success in introductory physics courses using a statistical method for establishing cutoff scores. *Learning and Teaching in Higher Education: Gulf Perspectives*, 14(1). <http://doi.org/10.18538/lthe.v14.n1.277>

